Article:
Walker, G. orcid.org/0000-0001-5022-4756 (2017) Visual Representations of Acoustic Data: A Survey and Suggestions. Research on Language & Social Interaction. ISSN 0835-1813

https://doi.org/10.1080/08351813.2017.1375802

This is the author's final version of an article to be published in *Research on Language and Social Interaction* (http://www.tandfonline.com/toc/hrls20/current), 2017, 50(4). http://dx.doi.org/10.1080/08351813.2017.1375802 .

Visual representations of acoustic data: a survey and suggestions

Abstract

Visual representations of acoustic data are becoming more common in Conversation Analysis (CA) and Interactional Linguistics (IL) research. This article provides a survey of visual representations of acoustic data in the journal *Research on Language and Social Interaction*. Shortcomings in their preparation and use are identified and discussed. Comparisons are made with visual representations prepared by expert phoneticians. Suggestions are made as to how visual representations could be prepared and used more effectively to support CA/IL researchers' claims and to allow readers to independently verify them.

1 Introduction

It is increasingly common for visual representations of acoustic data to be provided in order to convey information about what the researcher can hear in recordings of talk-in-interaction. Visual representations may also be used to illustrate measurements the researcher has taken. These visual representations include pitch traces, waveforms and spectrograms. Visual representations are an important resource for providing corroborative evidence for the researcher's claims as the reader will not normally have access to the original recordings. They can also allow the reader to independently verify those claims. How effectively visual representations serve these purposes depends on how researchers prepare and use them. As there is every reason to think that visual representations of acoustic data will become more common in CA/IL research, it is important to take stock of how they are being prepared and used.

Some articles by CA/IL researchers discuss the visual representation of acoustic information relating to particular phenomena (e.g. Couper-Kuhlen, 1996), but this is the first general survey of how visual representations of acoustic data are prepared and used in CA/IL research. This article provides a survey of how visual representations of acoustic data have been prepared and used in articles in *Research on Language and Social Interaction* (ROLSI). ROLSI is selected as an internationally recognised publication aimed at researchers in Conversation Analysis (CA) and Interactional Linguistics (IL).

The main aims of this article are to assist CA/IL researchers in producing (or improving) visual representations of acoustic data for use in their research by highlighting shortcomings in the visual representation of acoustic data and to propose some solutions. A further aim is to provide researchers with better access to

the information provided in visual representations of acoustic data prepared by others. It is not the aim of this article to show how phonetic analysis as such should be performed, but how some outcomes of that analysis can be given effective visual representation. Ogden (2009) is an excellent resource which could serve as an introduction to auditory and computer-based phonetic analysis for CA/IL researchers without experience of technical phonetic analysis: it describes auditory and computer-based techniques, and much of the book is based on conversational data. Walker (2013) discusses practical issues in the phonetic analysis of conversation, and gives an overview of findings which have emerged with regard to the role of phonetic details in the organization of conversation.

The article proceeds by first assembling the set of ROLSI articles containing visual representations of acoustic data (*Materials*). Aspects of these visual representations are evaluated for their effectiveness in providing corroborative evidence for the claims being made and allowing the reader to independently verify those claims (*Survey of visual representations of acoustic data).* Comparisons are made with relevant visual representations in articles written by expert phoneticians, some of which appear in specialist phonetics journals. Methods to help maximize the usefulness of visual representations of acoustic data are set out (*Methods in preparing visual representations*). The article ends with a summary of the main findings of the evaluation and the advice concerning the effective preparation and use of visual representations of acoustic data.

2 Materials

The online database *Web of Science* was searched for all articles in ROLSI. At the time of the search (January 5, 2017) titles were available from volume 28, issue 1 (1995) to volume 49, issue 4 (2016); abstracts accompanied articles from volume 32, issue 3 (1999). The search returned 375 items. The titles and abstracts of the returned items were then searched for terms relating to the phonetic design of talk. Each item returned in the search was verified to ensure that the term was being used to refer to speech production. In what follows *source* refers to an article containing one or more of the search terms; if a source contains a search term used to refer to speech production then this is called a *hit* for that term.[1] Table 1 shows the sources and hits, which of the sources include visual representations of acoustic data, which types of visual representations are included and how many figures appear in each source. There are 8 sources which include visual representations of acoustic data.

[1] Some search terms returned no hits: loudness, duration, rhythm, creak, tempo.

		phonetics term							visual representation			
ID	Source	prosod*	phonetic*	inton*	pitch	phonolog*	whisper	articulat*	pitch	waveform	spectrogram	number of figures
A	Mitchell (2001)				+							
B	Nakamura (2001)					+						
C	Wiggins (2002)			+								
D	Helasvuo, Laakso, and Sorjonen (2004)					+						
E	Hepburn (2004)				+		+					
F	Hellermann (2005)	+	+						+			3
G	Rendle-Short (2005)	+		+								
H	Hepburn and Potter (2007)						+					
I	Innes (2007)			+								
J	Betz and Golato (2008)	+			+							
K	Golato and Fagyal (2008)	+			+				+	+	+	2
L	Robinson and Kevoe-Feldman (2010)			+								
M	Wilkinson, Beeke, and Maxim (2010)	+										
N	Barth-Weingarten (2011)	+	+		+				+			3
O	Golato (2012)		+						+			3
P	Heritage (2012)	+										
Q	Pillet-Shore (2012)	+							+			8
R	MacMartin, Coe, and Adams (2014)	+										
S	T. Walker (2014a)	+	+	+	+							
T	T. Walker (2014b)		+									
U	Clayman and Raymond (2015)			+				+	+		+	6
V	Szczepek Reed (2015)		+						+	+		3
W	Stivers and Sidnell (2016)		+									
X	Szczepek Reed and Persson (2016)		+			+			+	+		2
total (24 sources)		10	8	6	6	3	2	1	8	3	2	30

Table 1 Phonetics terms and visual representations of acoustic data in *Research on Language and Social Interaction* article titles and/or abstracts, 1996–2016

The 8 sources which include visual representations of acoustic data contain a total of 30 figures, some of which are composite figures containing more than one kind of visual representation. The number of sources per year has stayed relatively steady since the first source up to the end of the study period.[2] This is shown in Figure 1. The figure also shows a steady increase in the proportion of sources which include visual representations. This rise corresponds with increased availability of digitally stored data and computer software capable of producing publication-quality graphics. At the end of the study period one-third of all sources contained visual representations; more than half of all sources since 2011 contain them. This increase in the use of visual representations is evidence of researchers' perception that they

[2] The average number of articles per year since 2001 is 20 (s.d. = 4).

are useful. As there is every reason to think that this trend will continue it is important to take stock of how visual representations are being prepared and used.

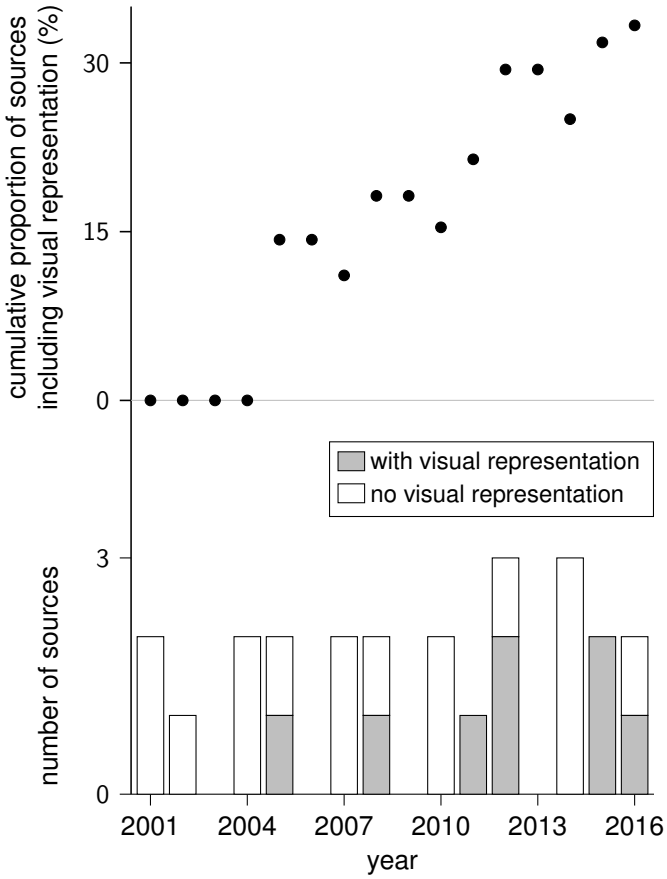


Figure 1 Sources which are hits for phonetics terms in ROLSI articles, 1995–2016. Bottom panel shows the number of sources per year which are hits, and the proportion of those sources which include visual representations of acoustic data; top panel shows the proportion of all sources which are hits for phonetics terms which include visual representations of acoustic data

A survey of visual representations is presented in the next section.

3 Survey of visual representations of acoustic data

This section reports on a survey of the visual representations in the sources in Table 1. The survey is concerned with the effectiveness of the visual representations in terms of providing corroborative evidence for the claims being made and allowing the reader to independently verify those claims. In other words, the survey considers how well the visual representations allow the reader to 'get at' the data, given that the reader will not generally have access to the original recordings. The survey is arranged around themes which arise from the consideration of the visual representations: the choice of the most effective visual representation, their interpretability, the awareness of relevant issues in the perception of sound, the use of visual representations in the text, the efficient use of space, the relevance and maximization of information conveyed, and presentational errors.

3.1 Choice of the most effective visual representation

Different kinds of information can be readily shown on visual representations: pitch, loudness, duration, and articulatory and phonatory quality can be shown via pitch traces, intensity traces, waveforms, spectrograms and other kinds of plots. However, the visual representation chosen to illustrate and support a claim may not always be the most effective one for the task.

Spectrograms are generally the most useful acoustic record for the visual representation of articulatory and phonatory details (for an accessible introduction to reading spectrograms, see Ogden, 2009.) However, only two sources include spectrograms: Barth-Weingarten, 2011 and Szczepek Reed, 2015 (source N and source V in Table 1, respectively; sources will be referred to with author-date labels and source letters on first mention, thereafter only source letters). Furthermore, claims may be made about articulatory and phonatory details yet they go without visual representation. For example, source N argues for the relevance of glottal closure at the end of German “ja” for turn-transition (no glottal closure) or turn-holding (glottal closure) (source N, p. 169-172). While pitch traces are used in support of claims concerning pitch, there is no visual representation provided to illustrate the presence or absence of glottal closure. Source V uses waveforms and pitch traces to illustrate the acoustic differences between glottal closure, creak phonation and voicing. An exemplar is shown in Figure 2.

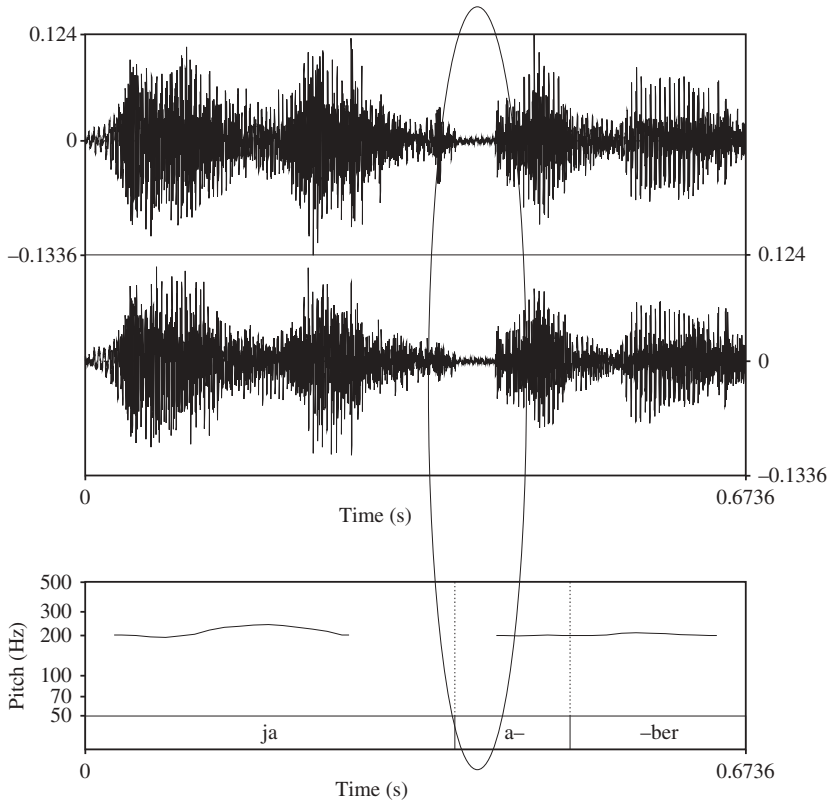


Figure 2 Labelled pitch trace and waveform from Szczepek Reed (2015, p. 35, Figure 2)

Creak and other types of non-modal phonation can cause problems for pitch detection algorithms and therefore can make pitch traces unreliable. Spectrograms

are generally more useful when considering phonation types and glottal activity. Ogden (2001), published in the *Journal of the International Phonetic Association*, provides a technical phonetic and sequential account of the turn-yielding and turn-holding function of creak and glottal stop in Finnish. Figure 3 is a visual representation of part of the utterance shown in Figure 1 of Ogden (2001), drawn to take into account some of the suggestions made in this article and to emphasise some of the relevant features. It illustrates the occurrence of creak at a point of possible turn-completion; Ogden (2001) provides a similar figure to illustrate glottal stop.

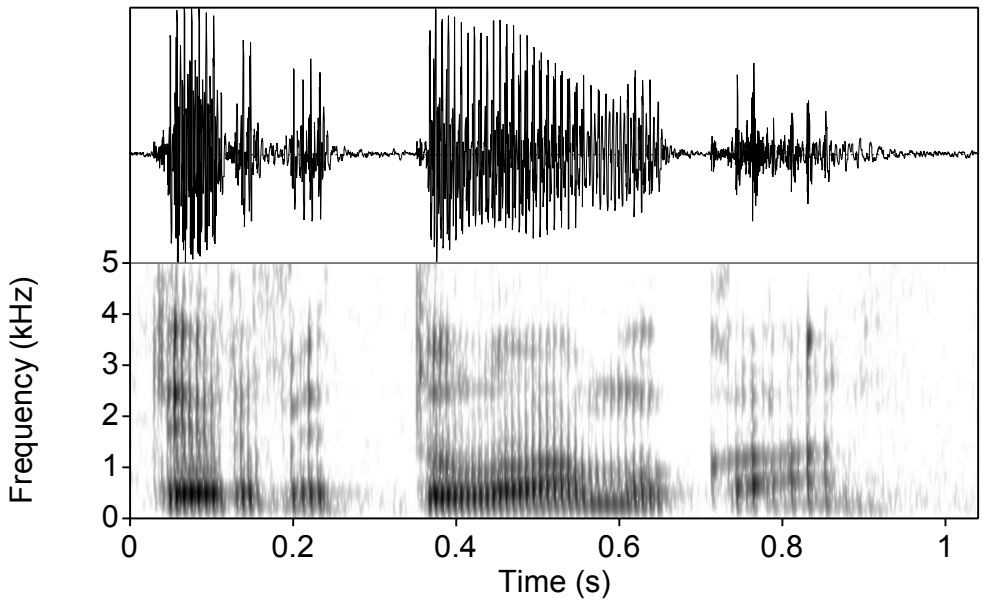
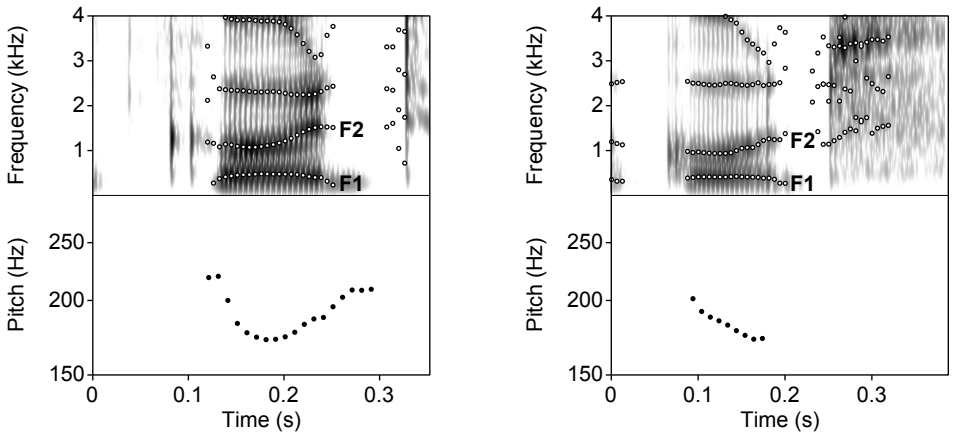


Figure 3 Waveform and spectrogram of an adult male producing Finnish "tervetuloa mukaan"

The occurrence and extent of creak is clear from the visual representation: the spacing and energy of striations in the spectrogram and pulses in the waveform are irregular between 0.7 s and 0.9 s (cf. the regularity of striations and pulses between 0.36 s and 0.65 s which reflect regular vocal fold vibration rather than creak).

Hellerman (2005) [source F] refers to the relevance of rhythm and timing to the organisation of talk interaction into segments, yet unlike pitch these features receive no visual representation (cf. Figures 7 and 8 in Ogden, 2013 in the *Journal of the International Phonetic Association* where rhythm is represented visually via labelled spectrograms and waveforms). Pillet-Shore (2012) [source Q] argues that as well as pitch and duration features, loudness and audible smiling are relevant to the display of stance. Pitch and duration features are reflected in the visual representations, while loudness and audible smiling are not. Loudness could have been shown via intensity traces. Since smiling causes formant frequencies, especially the second formant (F2), to rise (Barthel & Quené, 2015; Podesva, Callier, Voigt, & Jurafsky, 2015; Tartter, 1980) it may have been possible to use spectrograms and/or formant tracks in support of the claims about audible smiling by comparing formant frequencies during smiled speech with ones in comparable vowels from non-smiling speech. Spectrograms were used by Kohler (2008), published in *Phonetica*, for the visual representation of acoustic differences between speech produced with an audible smile and speech produced without an audible smile in German spontaneous

dialogue. Figure 4 is a visual representation of parts of the utterances shown in Figure 3b of Kohler (2008), drawn to take into account some of the suggestions made in this article and to emphasise some of the relevant features. Formant tracks have been overlaid on the spectrograms with labels placed at the end of the vowel portions of each token to identify the first and second formants (F1 and F2). Among other features discussed by Kohler, the figure shows that F1 and F2 have higher frequencies for the vowel [u:] when produced with speech-smile than when produced without. (Kohler states that F1 = 468 Hz and F2 = 1071 Hz for the speech-smile version; F1 = 409 Hz and F2 = 936 Hz for the non-speech-smile version.) The figure also shows that, as Kohler describes, the speech-smile version is produced with rising pitch whereas the non-speech-smile has falling pitch.



(a) with speech-smile (b) without speech-smile

Figure 4 Spectrogram and pitch trace of an adult female producing German "ja gut"

3.2 Interpretability of visual representations

It may be unnecessarily difficult to interpret relevant information from visual representations and in some cases impossible. In other words, a reader inspecting a visual representation may find its contents to be at odds with what is to be expected from other visual representations in the source, and/or from facts about speech production. There may be inadequate explanation of how data-points were arrived at. This section discusses several sources where such problems of interpretability arise, and describes ways in which these problems might have been avoided.

Source Q provides an account of the prosodic design of greetings. Eight pitch traces with accompanying word labels are provided to support and illustrate claims concerning pitch characteristics and duration. An example is shown in Figure 5.

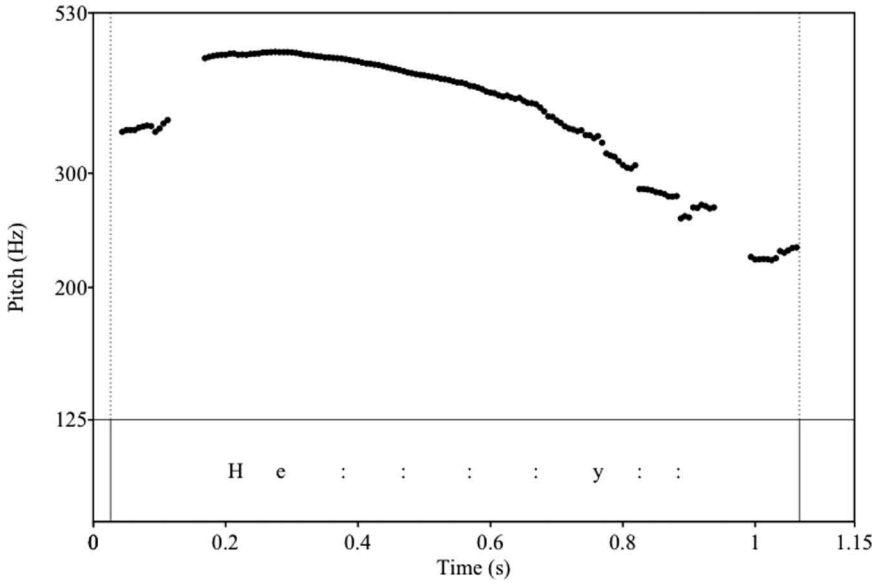


Figure 5 Labelled pitch trace from Pillet-Shore (2012, p. 379, Figure 1)

It would seem reasonable to assume that in each visual representation the boundaries at the start and end of the labelled portion mark the start and end of detectable speech production and that an empty interval on the text tier means no speech can be detected in that portion. However, initial boundaries do not seem to be consistently placed at the start of detectable speech in this source.

Pitch analysis deals with the fundamental frequency of speech sounds. Voiceless sounds are aperiodic and do not have a fundamental frequency, therefore an accurate pitch analysis will not include values for voiceless portions. It seems that at least some of the initial boundaries in source Q are placed after the onset of detectable speech. All but one of the greetings are versions of “hi” or “hey”: words which ordinarily begin, in turn-initial position, with a voiceless glottal fricative [h]. In most of the pitch traces the boundary at the start of the label for the greeting is placed at the point where the dots indicating pitch values (*pitch-dots*) begin, or just before it as in Figure 5. In Figures 6 and 7 the initial boundaries are placed where pitch-dots begin.

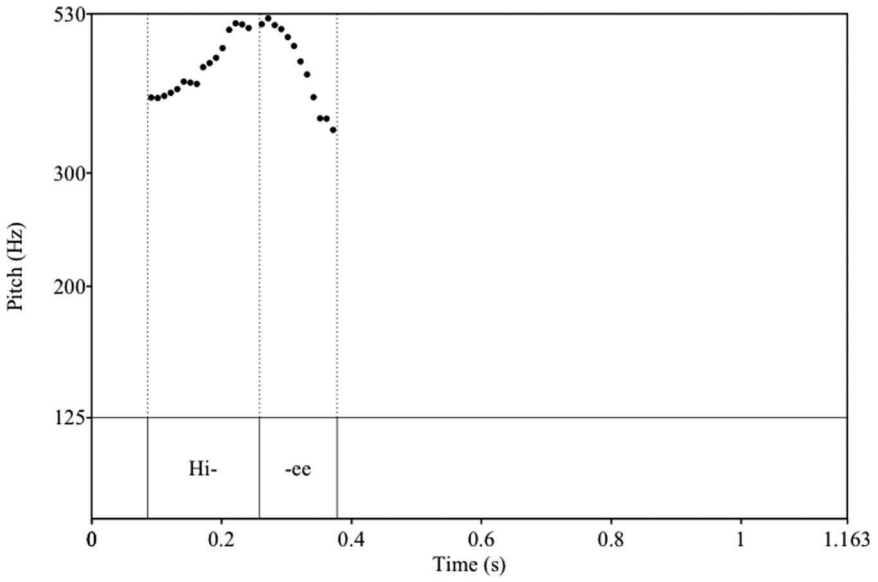


Figure 6 Labelled pitch trace from Pillet-Shore (2012, p. 379, Figure 2)

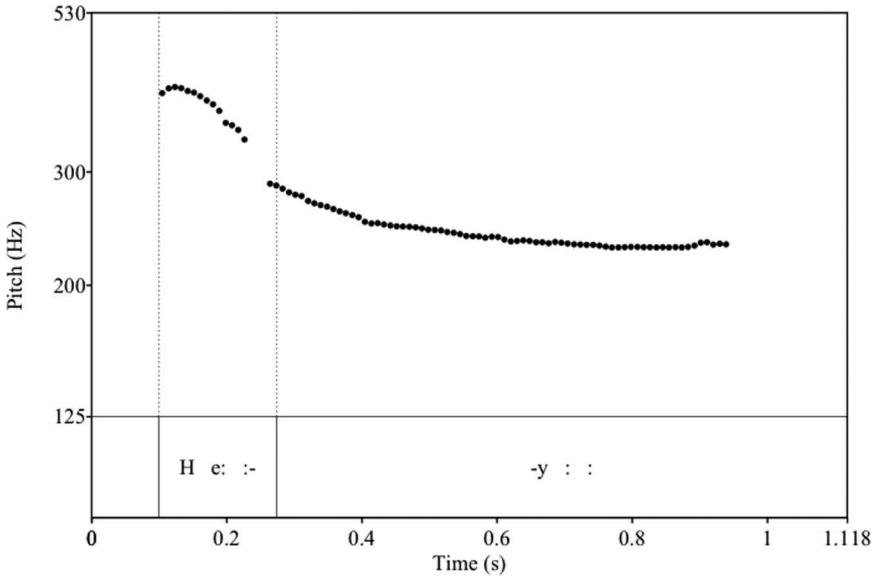


Figure 7 Labelled pitch trace from Pillet-Shore (2012, p. 381, Figure 3)

A gap between the initial boundary and the beginning of the pitch-dots would have been expected given the initial voiceless sounds indicated in the transcriptions. To add to the reader's uncertainty over the placement of the boundaries in source Q, in some cases there is a gap between the initial boundary and the beginning of the pitch-dots which suggests the boundaries may have been placed at the onset of detectable speech.

It is claimed that "Paula substantially lengthens (i.e., sound-stretches) her production of the greeting term 'Hey' to 1.15 s in duration" (source Q, p. 378) and that this is evident from Figure 5. However, the only way to interpret Figure 5 as showing that "Hey" has a duration of 1.15 s is by looking at the total duration of the material represented in the figure. This would make the boundaries redundant and undermine the assumption that an unlabelled interval on the text tier means no speech can be detected. But this principle of ignoring boundaries cannot be consistently applied to all figures.

It is not possible to verify the author's claims concerning the duration of tokens from the visual representations presented in source Q due to the way they are prepared and discussed. It is claimed in the text that the token in Figure 6 has a duration of 0.29 s (source Q, p. 380). This duration seems to refer to the time from the first boundary (at ≈ 0.08 s) to the last (at ≈ 0.38 s).[3] This means that to arrive at the author's conclusions concerning the duration of the tokens the reader has to pay attention to the initial and final boundaries in one case (Figure 6) but ignore them in another (Figure 5). In Figure 7 there is a boundary at the start of the greeting but no boundary at the end. It seems reasonable to assume that the token extends to the right-hand edge of the visual representation. This token is reported in the text as "totaling 0.85 s in duration" (source Q, p. 381). A duration close to the author's claim

[3] Estimates from visual representations have been arrived at by taking measurements from the PDF versions of the source using Preview on Mac OS(X). Estimates are identified by ≈ in the text.

of 0.85 s can be reached if we assume that the token starts with the first boundary (≈ 0.1 s), and ends with the final pitch-dot (≈ 0.94 s). This is not how other visual representations had to be interpreted.

The problems of interpretability in source Q are unfortunate as duration is claimed to be an important aspect of the design of the greeting tokens and a means for conveying the greeter's stance towards their relationship with the person being greeted. Duration information could have been straightforwardly shown by showing no more and no less than all detectable speech associated with each greeting in each visual representation: the reader could then have taken the amount of time shown in each visual representation to be the duration of the token. If there was a need to reflect relative duration visually, then each visual representation could have been scaled so that the amount of horizontal space taken up by the visual representation corresponded to the duration of each greeting. The boundaries should have had a consistent meaning across all visual representations, for example marking the onset and offset of detectible speech. Problems in interpretation are compounded by the decision not to include waveforms and/or spectrograms: correctly prepared these would have allowed the reader to independently identify the likely start and end of detectible speech.

Kaimaki (2011), published in *Journal of Pragmatics*, provides a technical phonetic and sequential analysis of English call openings and avoids several of the shortcomings evident in the visual representations in Source Q. Kaimaki describes her work as forming "part of a larger research programme which seeks to provide accounts of phonetic variability and phonological organisation by reference to the sequential organisation of talk" (Kaimaki, 2011, p. 2147). One kind of visual representation in Kaimaki (2011) is shown in Figure 8, and can be compared with Figures 5–7 above from source Q.

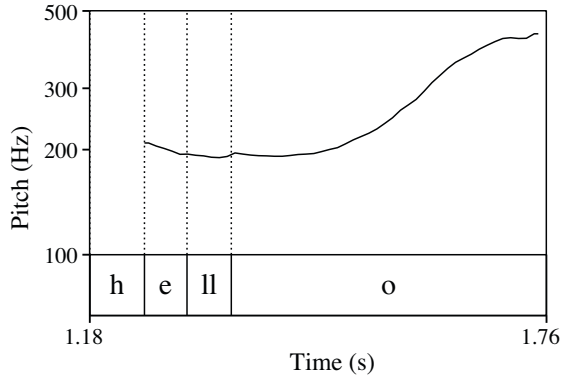


Figure 8 Labelled pitch trace from Kaimaki (2011, p. 2134, Figure 1)[4]

In Figure 8 the boundaries identify the margins of individual speech sounds. The initial sound which would be expected to be voiceless does not have pitch values, and only those portions where there is detectible speech seem to have been included. Figure 8 could have been enhanced by the inclusion of a waveform or a spectrogram (or both). This would have allowed the reader to verify the labeling and to check whether portions without pitch values are voiceless. Figure 9 shows the same portion of the recording as Figure 8, drawn using the same axes, and including a waveform and spectrogram. The waveform is aperiodic during the initial sound and

[4] Reprinted from Journal of Pragmatics, 43 (8), Marianna Kaimaki, Transition relevance and the phonetic design of English call openings, 2130-2147, Copyright (2011), with permission from Elsevier.

there are no striations in the spectrogram, both of which support the claim that the initial sound is voiceless and explain the absence of pitch values during that portion. (A waveform is said to be periodic where the pattern in the waveform repeats, these repetitions corresponding to vibrations of the vocal folds: see Ogden, 2009, p. 30–32.) It can be seen that the boundaries have been placed where changes in the waveform and spectrogram suggest changes in articulation and phonation.

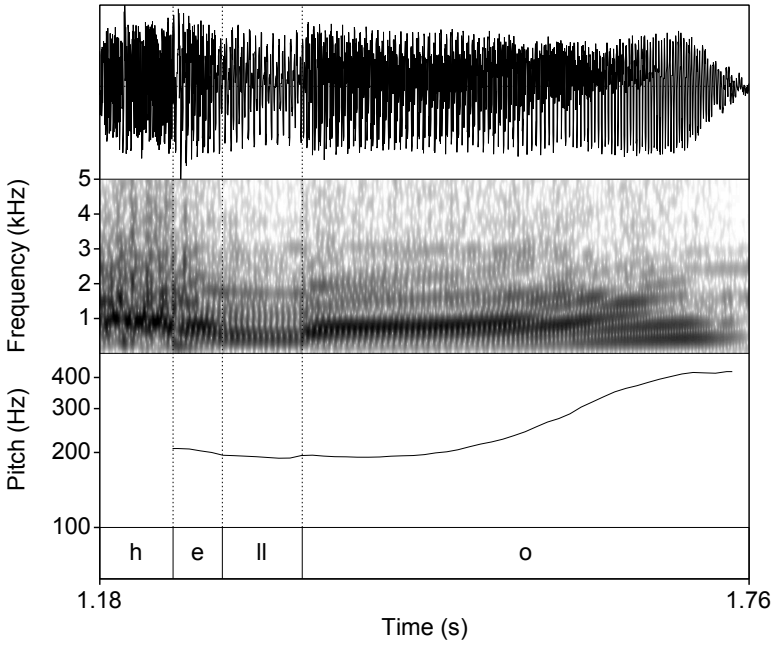


Figure 9 Labelled waveform, spectrogram and pitch trace of the same utterance as Figure 8

Source F also provides visual representations of pitch. They are provided as part of an account of sequential, syntactic and prosodic features of three-part sequences of classroom talk. All three representations take the form of Figure 10. In the figure four utterances are each represented by a data-point, with pitch represented on the y-axis.

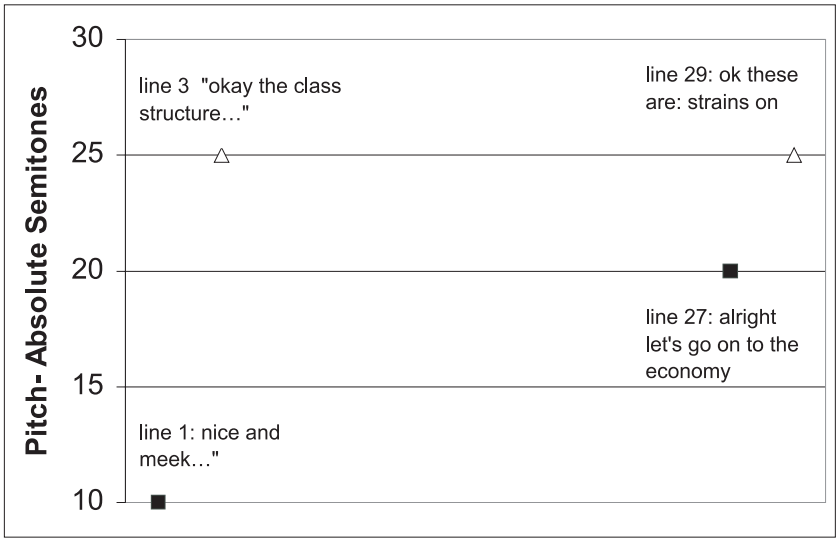


Figure 10 Representation of pitch from Hellermann (2005, p. 111, Figure 1)

There is no straightforward account of how these utterances came to be represented by single data-points. An utterance could conceivably be reduced to a data-point representing (for example) the mean of the utterance, the median, the minimum, the maximum, the beginning, the end, the height of the pitch peak, and so on. The choice is consequential as these measures will typically be rather different from one another. It is stated in the text that the difference between the data-points for line 1 and line 3 equates to a pitch reset of 100% for this segment (source F, p. 112). Pitch reset is determined by taking into account the difference in pitch relative to the pitch range of the interactional segment. The pitch range is defined as the "difference between the highest and lowest pitch" (source F, p. 127, footnote 7). The obvious way to interpret "highest and lowest pitch" is as referring to maximum and minimum pitch values for the utterances. However, this interpretation cannot be applied to all visual representations of pitch in source F: the data-points in Figure 2 of source F are described as representing pitch peaks (source F, p. 115). Neither the description of Figure 10 nor its caption ("Pitch reset at the start and close of the talk from Excerpt 1", source F, p. 111) suggest the reader should interpret the data-points on Figure 10 as representing pitch peaks. Furthermore, interpreting the data-points in Figure 10 as representing pitch peaks seems to contradict the supplied definition of pitch range: even if the data-point for line 1 represents a low pitch peak then there will be a lower minimum pitch value in the utterance, in which case pitch range cannot be the difference between a maximum and minimum value.

The situation arising in source F is similar to that which arose in source Q: it is unclear how equivalent visual representations are to be interpreted, and they may have to be interpreted in different ways. In some cases it is not possible for the reader to interpret visual representations confidently and consistently; in other cases they are made unnecessarily difficult to interpret. Source N analyses features of "ja" and "jaja" in German. Three labelled pitch traces of tokens are provided to illustrate patterns in the pitch characteristics of the tokens, as shown in Figures 11 and 12.

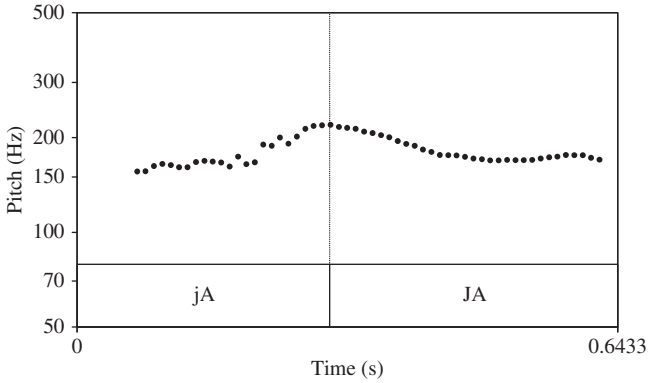


Figure 11 Labelled pitch trace from Barth-Weingarten (2011, p. 167, Figure 1)

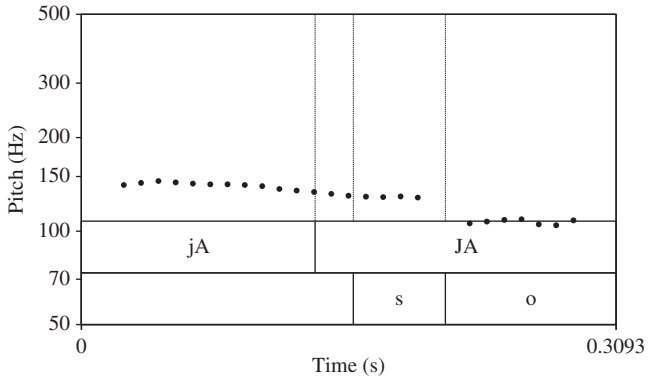


Figure 12 Labelled pitch trace from Barth-Weingarten (2011, p. 176, Figure 3)

The horizontal line above the text tier in Figures 11 and 12 gives an immediate visual impression of some kind of lower limit on the plot. (This is the function of the horizontal line above labels in Figures 5–7, for example.) In Figure 12 the horizontal line above the labels is at a higher frequency than in Figure 11. There seems to be no reason for this other than to accommodate two tiers of labels. This shows that although the horizontal line above the labels may give the visual impression of some kind of lower limit (for example of the plot area, or the speaker's pitch range) it is in fact decorative and simply sets off the labels from the main plot. Part of the analytic account in source N concerns the placement of utterances in the speaker's pitch range, so strategic and meaningful placement of the horizontal line at the bottom of the speaker's pitch range would have been helpful.

In summary, a reader may find it impossible to fully and confidently interpret visual representations of acoustic data in a source. There may be inconsistencies in presentation within and/or across visual representations of acoustic data. The contents of visual representations may be at odds with what the reader would expect based on facts about speech production. There may be inadequate explanation of how data-points were arrived at. Visual representations of acoustic data may be unnecessarily difficult to interpret. These limitations may make it unnecessarily difficult or impossible for the reader to use the visual representations to verify the claims being made.

3.3 Perceptual awareness in visual representations

Visual representations should reflect what is known about the perception of sound as far as is possible and practical. Visual representations in several sources take into account the non-linear perception of frequency by presenting pitch traces on a logarithmic scale (source N, source U [Clayman & Raymond, 2015], source Q, source V). Data-points are plotted in semitones in source F which also takes into account the non-linear perception of frequency. The pitch traces in Golato & Fagyel (2008) [source K] and Golato (2012) [source O] are plotted on a linear scale: see Figures 13 and 14. A logarithmic (non-linear) scale better reflects the non-linear auditory perception of frequency. (This is discussed in more detail later.) This is of particular relevance to the visual representations in source O: a linear scale may make pitch peaks look more extreme than a non-linear scale.

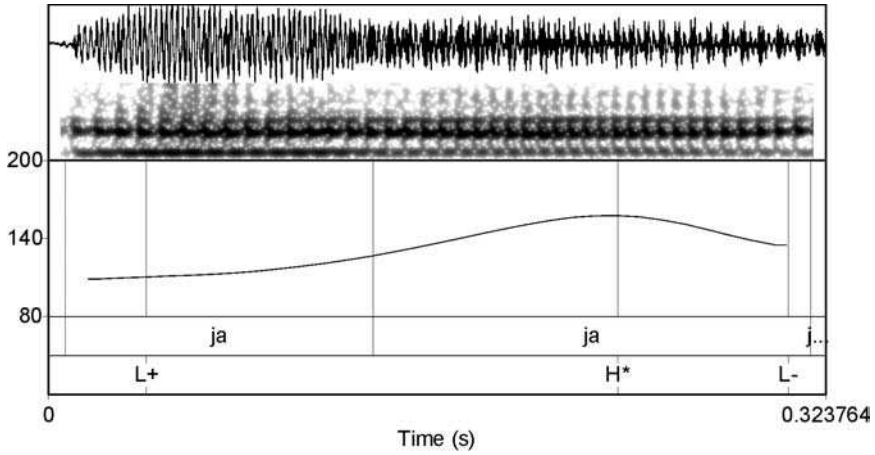


Figure 13 Labelled waveform, spectrogram and pitch trace from Golato and Fagyal (2008, p. 252, Figure 2)

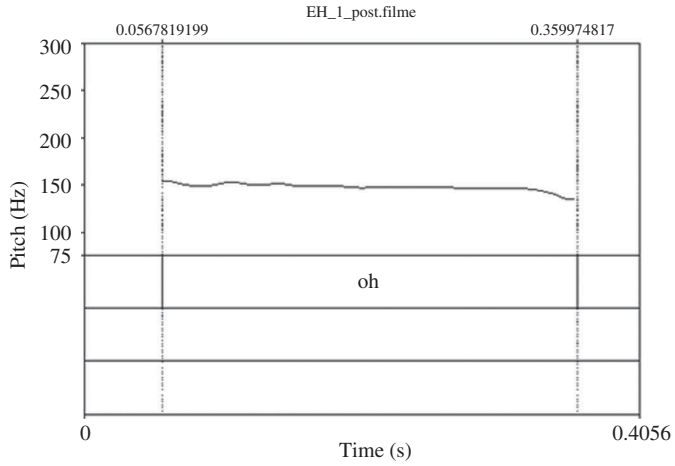


Figure 14 Labelled pitch trace from Golato (2012, p. 263, Figure 2)

The same, essentially arbitrary, y-axis scale is used across all pitch traces within several sources: 50–500 Hertz (Hz) in source N, source O and source V; 75–500 Hz in source U; 75–600 Hz in Szcepek Reed and Persson, 2016 [source X]). Preparing a y-axis which reflects the speaker's pitch range provides a better visual reflection of the placement of the talk in the speaker's pitch range, and how much of the speaker's range the pitch movements cover (see section 4.2 on practical aspects of adjusting the y-axis of a pitch trace to represent the speaker's pitch range). It therefore becomes possible to visually compare tokens: it is well known that what counts as 'high' for one speaker may count as 'low' for another (for example) due to factors such as gender, age and body size (Hollien, Hollien, & de Jong, 1997; Kreiman & Sidtis, 2011; Nishio & Niimi, 2008). Furthermore, research in CA/IL has shown that the placement of talk in the speaker's range is interactionally important (Couper-Kuhlen, 1996; Local, 2005). A y-axis which reflected the speaker's pitch range would have been especially helpful in source N and source O, where the text refers specifically to the placement of talk in the speaker's range: something which cannot be evaluated from the visual representations in those sources.

The visual representations of pitch in source F could have conveyed more information if they had represented the pitch range of the interactional segment. The

y-axis on Figure 10 could have run from the bottom of the pitch range for the segment to the top. This would have provided a better visual reflection of the placement of utterance values in the pitch range of the interactional segment rather than distance from an arbitrary and unspecified reference value. It is especially important to specify the reference value for semitone calculations as there are several reference values in use: Praat (Boersma & Weenink, 2017) provides commands calculating Hertz values in semitones relative to 1 Hz, 100 Hz, 200 Hz, and 440 Hz; Fletcher (1934) proposes a reference value of 16.35 Hz. Pitch range is relevant to the author's argument as it is a way of determining the occurrence of pitch reset (see source F, p. 127, footnote 7).

3.4 Use of visual representations in the text

If visual representations are presented in order to corroborate the researcher's claims and to allow the reader to independently verify them then the visual representations could be expected to be tied closely to the text. This is not always the case. Source X provides two visual representations to illustrate the difference between glottalized and joined-up word boundaries, one of which is shown in Figure 15.

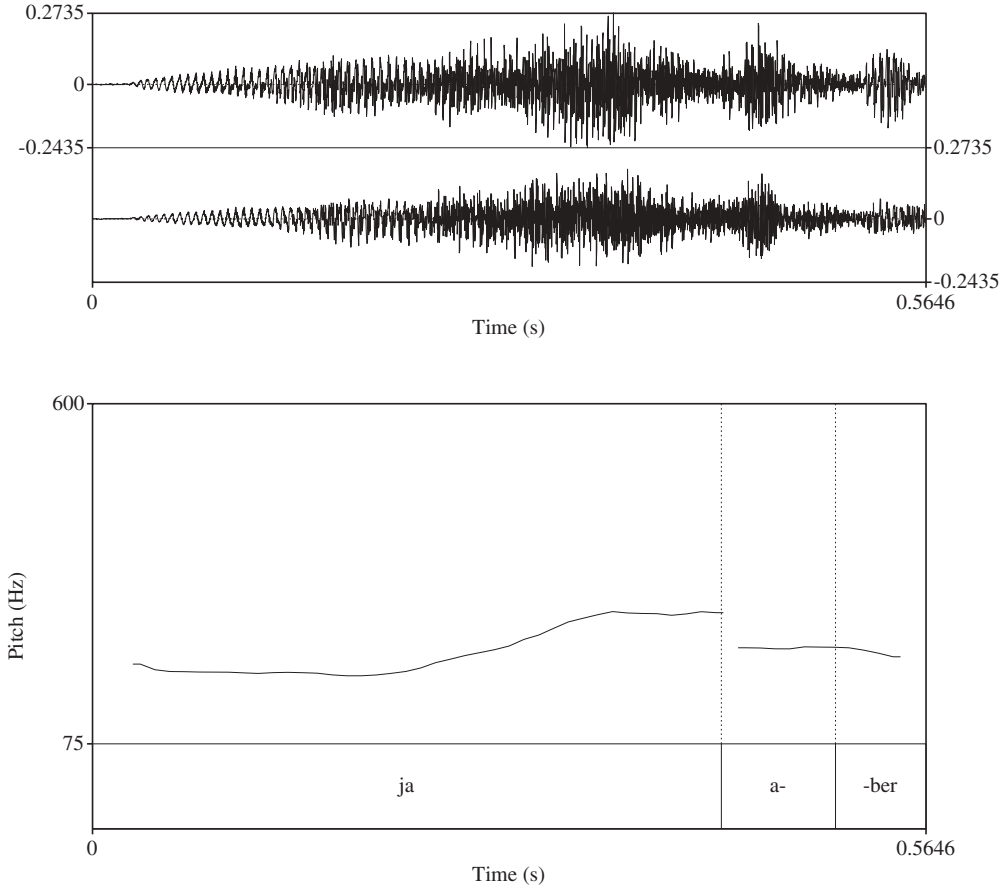


Figure 15 Labelled waveform and pitch trace from Szczepek Reed and Persson (2016, p. 133, Figure 2)

However, there is no discussion of these visual representations in the text. As a result it is unclear how they demonstrate the two different phonetic possibilities and

an opportunity missed to use visual representations as corroborative evidence for the authors' claims. Most of the spectrograms in source U are not discussed at all even though they might have been used to support the points being made such as reference to a break in voicing and no glottal stop at the start of the turn-extension. Two visual representations are provided in source K, each of which contains a labelled waveform, spectrogram and pitch trace: see Figure 13.

The text does not refer to the waveforms or spectrograms when they might have been used to support some of the points made. For example, both records show (in different ways) the continuation of vocal fold vibration from one "ja" token to the next, which provides an empirical basis for the claim that the second "ja" may be produced "in immediate succession" (source O, pp. 9 and 13).

While visual representations may be underused in the text, in some cases the text may lead a reader to think that they show something different from what is actually shown. Source F states that "[t]he prosodic packaging of the activity segment from Excerpt 1 is shown in Figure 1 [Figure 10 in this article]" (source F, p. 111). However, the visual representation only shows pitch information: no other aspects of prosody are represented. While there is variation in which features researchers are referring to when they use the term *prosody* (see Peppé, 2009), a common use is to refer to features not just of pitch, loudness and duration (rhythm and rate). Source N implies that pitch range is reflected by the layout of the visual representations: "the pitch range covered is very small; it starts and ends around the middle of the speaker's range (see Figure 3 [Figure 12 in this article])" (source N, p. 175). However, the visual representation conveys no information at all about pitch range, being presented on the same scale as all other pitch traces in the source. Source O suggests that details are shown in a visual representation which are not: "A number of *oh*s feature a bell-shaped intonation contour typically spoken louder than the surrounding talk, as exemplified in Figure 1, which depicts Excerpt 12" (Source O, p. 262). However, the visual representation, which has the same format as Figure 11, only contains a pitch trace and does not convey any information about loudness. Source Q says that "[t]o facilitate comparison, the window size of Figure 7 is on par with other figures in this article" (source Q, p. 391). However, there is considerable variation in the amount of material shown, from 0.905 s (Figure 4 of source Q) to 1.807 s (Figure 8 of source Q): visual comparison of the representations is therefore not as straightforward as the author's statement suggests.

3.5 Efficient use of space

Some visual representations do not use space efficiently to convey relevant information. Space is wasted in the visual representations in source O in two ways, both exemplified by Figure 14. The parts of the visual representation before and after the token (indicated by boundaries) and the two unused tiers for word labels do not convey anything to the reader. These areas account for ≈ 32% of the space taken up by Figure 14, ≈ 54% of Figure 1 of source O and ≈ 46% of Figure 3.

Visual representations in source V and source X contain waveforms and pitch traces in separate boxes, each with identical labels on the x-axis: see Figures 12 and 15. This means that ≈ 13% of the area of the visual representation conveys nothing at all to the reader. The area taken up by the visual representations could have been reduced by plotting the pitch and waveform adjacent to one another with a single labelled x-axis: see Figure 16.

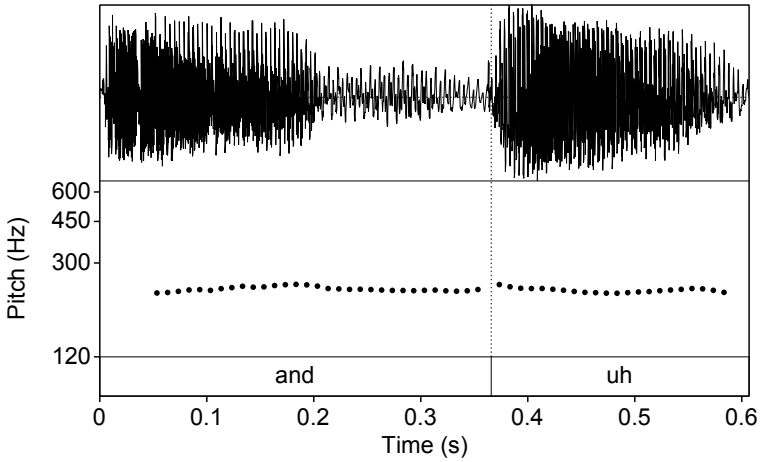


Figure 16 Labelled pitch trace and waveform of an adult female's speech; the y-axis represents the speaker's normal speaking range[5]

This presentation also emphasises the relatedness of the pitch trace and waveform (the former is derived from the latter) rather than making them appear visually as two separate entities.

3.6 Conveying information by visual representations

Some visual representations convey information which is irrelevant to the tasks of providing corroborative evidence in support of the claims being made, and allowing the reader to independently verify those claims. There are other sources where visual representations could convey more information with modest modification.

In a survey of data-graphics Tufte (2001) defines *data-ink* as "the non-erasable core of a graphic, the non-redundant ink arranged in response to the numbers represented" (Tufte, 2001, p. 93). Tufte arrives at a *data-ink ratio* for graphics by dividing, literally or figuratively, the data-ink by the amount of ink used to print the graphic: the higher the data-ink ratio, the higher the proportion of ink being used convey data-information. It is to be expected that researchers would look to maximise the data-ink ratio. This is not simply a matter of aesthetics or saving ink: maximising the data-ink ratio maximises the amount of relevant information conveyed by a visual representation and minimises distractions from that information.

Even without precise calculations, it is clear that the data-ink ratio is reduced in several of the sources by having unnecessarily long final labels on the x-axis. For example, in Figure 6 and 7 from source N final times are stated to 4 decimal places: the ends of the tokens are being identified to the nearest ten-thousandth of a second. Even contemporary laboratory-based studies in phonetics usually work at a level no more precise than thousandths of a second (milliseconds). The same level of detail is apparent on x-axes in other sources (source O, source V, source X). Source K gives final times to 6 decimal places (millionths of a second). Such detail on a visual representation is not required to provide corroborative evidence in support of the claims being made, nor are they required in order to allow the reader to

[5] Reprinted from Journal of Pragmatics, 57, Rasmus Persson, Transition relevance and the phonetic design of English call openings, 19-38, Copyright (2013), with permission from Elsevier.

independently verify those claims.

The highest level of apparent precision is in source O. In Figure 14 (one of three equivalent visual representations in source O) the label accompanying the final boundary is given to 9 decimal places (billionths of a second), while the label accompanying the initial boundary is given to 10 decimal places (ten-billionths of a second). The labels suggest the author is trying to convey something about the duration of the tokens but the length of the labels makes determining the duration unnecessarily difficult: the reader needs to subtract 0.0567819199 from 0.359974817 to give 0.3031928971 (seconds). Instead of labelling boundaries a label saying “303 ms” could have been provided and the labels accompanying the boundaries removed. This would have increased the data-ink ratio: across the three visual representations the number of characters conveying information about the duration of the tokens would have gone down from 67 to 15, and analytically important information would have been available with the minimum amount of effort on the part of the reader. An even more efficient way to convey information about the duration of the tokens would have been to show only the token without the preceding and following silence and with a sensibly rounded label at the end of the x-axis (e.g. “0.3” in the case of Figure 14).

In many cases shorter labels on visual representations would convey just as much useful information to the reader and would increase the proportion of data-ink. There are other ways the data-ink ratio could be increased in some sources. Two visual representations of the type in Figure 15 appear in source X; three equivalent visual representations are found in source V. It is evident from the presentation of two different waveforms that the data have been recorded in stereo. However, since the differences are very slight and irrelevant to the authors’ arguments, the waveform of a single recording channel could have been presented. This would have increased the data-ink ratio without obscuring any relevant details and reduced the size of each figure by ≈ 15%.

3.7 Other issues in presentation

Some visual representations show basic presentational errors. Some of these errors make confident interpretation of visual representation impossible. For example, the visual representations in source X only include labels at the top and bottom of the y-axis (see Figure 15). This means the reader does not know if the pitch traces are being shown on a linear or a logarithmic scale. Figure 2 of source K (reproduced here as Figure 13) does not provide a y-axis title so the reader is left to assume that the units on the y-axis are Hertz. There are no axis labels for the spectrogram so the reader does not know which part of the acoustic spectrum is being presented. The title and labels are absent from both axes in Figure 3 of source O, making it impossible for the reader to tell anything about pitch height or span from the visual representation. Figure 1 of source K appears to be ‘bitmapped’ as a result of post-processing of the graphic. This has the effect of making the figure blocky and difficult to read; the pitch trace may also have been affected as it is much more jagged than the pitch trace in Figure 2 (reproduced here as Figure 13). Any effect on the pitch trace is especially unfortunate as the visual representations are there primarily to illustrate pitch features.

3.8 Summary

This section has surveyed some characteristics of visual representations of acoustic data in ROLSI. The next section sets out in more detail some more technical factors which could usefully be taken into account in the preparation of visual representations. To round off the discussion of shortcomings a visual representation with much to commend it as a means for supporting the author's claims and allowing the reader to independently verify them will be considered. The visual representation is from Persson (2013) (published in *Journal of Pragmatics*) and is reproduced in Figure 16. It effectively conveys information which several of the visual representations in ROLSI could, or should, have conveyed.

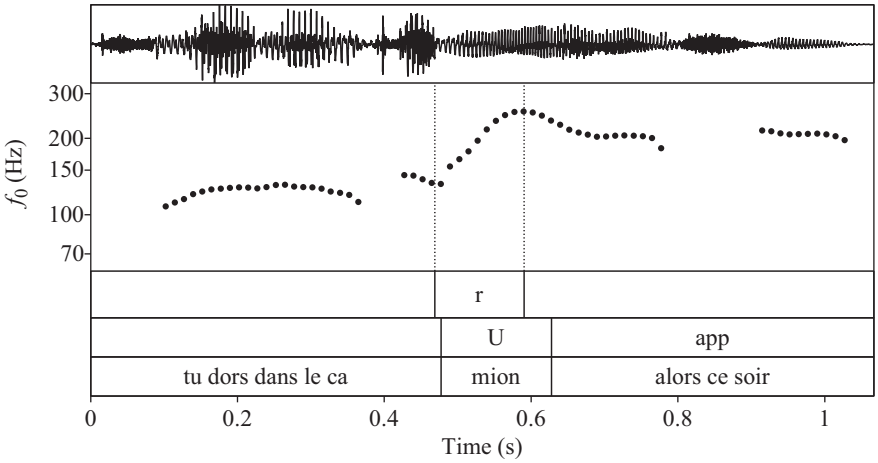


Figure 16. Labelled waveform and pitch trace from Persson (2013, p. 23, Figure 1)

Persson (2013) presents an account of intonational and sequential characteristics of French formulations (very roughly, summaries of a co-participant's talk). Figure 16 is intended to illustrate an intonational pattern (*final rise*) which the author describes. The pitch trace is plotted on a logarithmic scale representing the speaker's pitch range, the importance of which is discussed in the next section. The rise, demarcated by vertical lines and identified by the label 'r', can be easily identified in the figure. The author describes the rise as having its peak on the ultima (final accented syllable of a TCU) which is clearly labeled ('U'). It is possible to estimate the size of the rise from the figure as ≈ 10.6 ST: the lowest (first) pitch-dot in the rise is ≈ 140 Hz, and the highest (last) pitch-dot is ≈ 270 Hz. This difference can be expressed in semitones (ST) using a calculation based on Reetz & Jongman (2009, p. 243): 12×(*log*2(270/140)) = 11.4 ST. (The semitone scale is discussed in more detail later.) This rise can therefore be verified as within the limits on the phenomenon described by the author. The waveform allows the reader to confirm that pitch-dots appear only wherever the waveform is periodic (i.e. during voiced portions) and nowhere else. A tier of orthographic labels provides the reader with an indication of what is being said and the axes have appropriate labels.

4 Methods in preparing visual representations

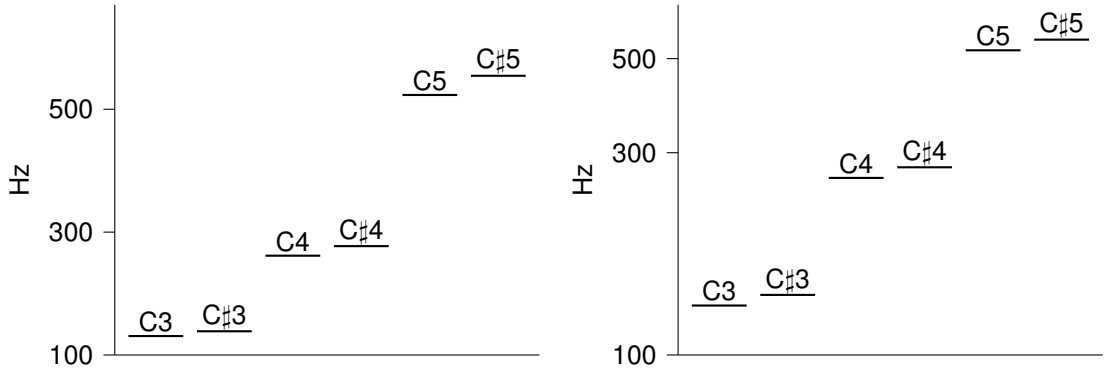
This section discusses methods to help maximise the usefulness to researchers in CA/IL of visual representations of acoustic data. These methods concern the plotting

of pitch traces on logarithmic scales, plotting pitch traces relative to the speaker's range, and the adjustment and use of spectrograms. The methods have been selected in light of the foregoing survey. The suggestions are intended to help researchers use visual representations to support their claims and allow readers to independently verify them.

4.1 Logarithmic and semitone scales for pitch

Many visual representations plot pitch values on a non-linear (logarithmic) scale. Since not all pitch traces in the sources are presented in this way, this section will explain why this is generally the most appropriate way to present pitch data.

The frequency response of the auditory system is not linear: humans are generally more sensitive to changes at lower frequencies (Johnson, 2003, pp. 88-90; Baken & Orlikoff, 2000, p. 148). The keys on a piano give a good way to understand this. There is a difference in pitch between the key C3 (twelve keys to the left of middle C) and the key immediately to the right of it (C♯3). This perceptual difference is the same as the difference between C4 (middle C) and the key to the right of it (C♯4). The perceptual difference between C5 (twelve keys above middle C) and C♯5 is the same as the difference within the other two pairs. The perceptual difference between the notes in each pair is the same. However, the difference in fundamental frequency of the notes produced by the keys is larger—exactly double—each time. Figure 17 shows how the fundamental frequencies of these sounds look when plotted on a linear scale (Figure 17a) and on a logarithmic scale (Figure 17b).



(a) Hertz on a linear scale (b) Hertz on a logarithmic scale

Figure 17 Selected musical notes plotted on different scales; horizontal lines indicate fundamental frequency of the notes

On the linear scale (Figure 17a) the vertical distance between the notes within each pair is greater for the second pair than the first, and greater again for the third. Also, the vertical distance between the second and third pair of notes is greater than between the first and second pair. However, this is not at all how humans perceive sounds: the difference within each pair is the same, as is the difference between the pairs. This is captured visually in the logarithmic plot (Figure 17b) where the vertical space within each pair is the same, as is the vertical space between the pairs. Plotting pitch on a logarithmic scale therefore provides a better visual reflection of our perception than the linear scale.

The non-linear frequency response of the auditory system means it often makes sense to express differences between two frequencies in semitones. Figure 18

shows the relationship between frequency in Hertz and pitch in semitones.

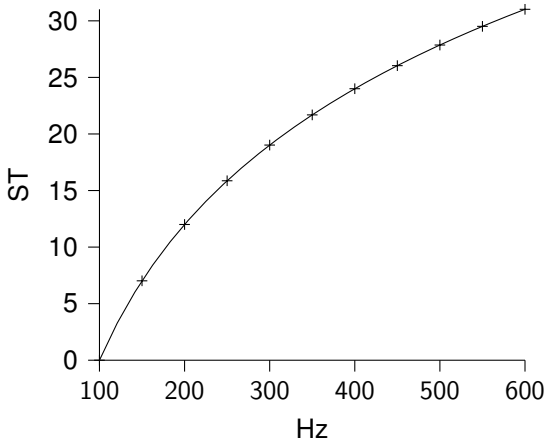


Figure 18 Hertz plotted on a linear scale against semitones calculated relative to 100 Hz; + is placed at 50 Hz intervals

The same change in frequency corresponds to a larger difference in pitch at lower frequencies than at higher ones. For example, the difference between a sound with a fundamental frequency of 100 Hz and another with a fundamental frequency of 200 Hz is 12 ST (or 1 octave); the difference between a sound with a fundamental frequency of 300 Hz and another with a fundamental frequency of 400 Hz—so, the same 100 Hz difference—is 4.98 ST. To return to the piano, the difference in frequency between C3 and C♯3 on a piano is 7.78 Hz, between C4 and C♯4 it is 15.55 Hz, and between C5 and C♯5 it is 31.12 Hz (Pierce, 1992, p. 19). However, the difference within each pair is always 1 ST. Pitch traces prepared in semitones allow the reader to compare data-points in terms of perceptual similarity and difference. Figure 18 shows a labelled pitch trace, plotted in semitones relative to the speaker's baseline. (A plot of this portion of talk on a logarithmic scale can be found in source U, p. 395.) This allows the reader to estimate pitch values in perceptually relevant units.

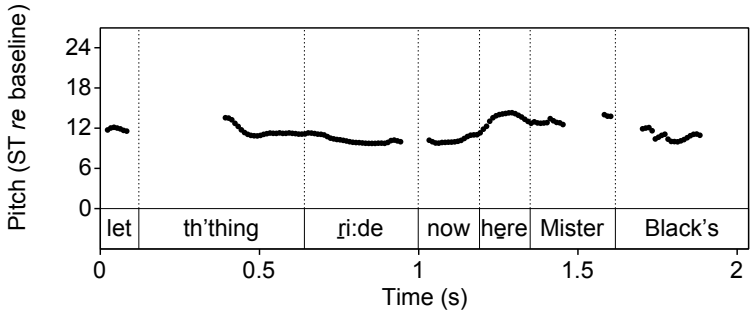


Figure 19 Labelled waveform and pitch trace of an adult female's speech; the y-axis represents the speaker's normal speaking range

The normalisation offered by the semitone scale allows for perceptually relevant comparison both within and between speakers: pitch values within one speaker's talk can be compared, and the placement of talk in the pitch range of multiple speakers can be represented.

4.2 Scaling to the speaker's pitch range

Adjusting the y-axis of a pitch trace to represent the speaker's pitch range allows the reader to see where an utterance is placed in the speaker's range and estimate how much of the speaker's pitch range the utterance covers (see Figure 19). Since only one of the sources did this (source Q), this section will discuss how the speaker's pitch range can be estimated in order to prepare such a visual representation.

In order to provide a visual reflection of the speaker's pitch range the top and bottom of the speaker's pitch range needs to be estimated. It may also be helpful to estimate the middle of the speaker's range.

A convention adopted in previous CA/IL research has been to consider 1 min of material. While it may be possible to gain reasonable estimates of the middle of the range from a smaller amount of material (see Horri, 1975, Lennes et al. 2015), it seems appropriate to err on the side of caution and consider more material: the more material considered, the more likely it is that the speaker will produce speech near the bottom and top of their range. To eliminate unreliable pitch measures and to gain a reasonable estimate of a speaker's pitch range, Praat can be used for visual and auditory comparison of the pitch traces with the original audio. (Praat is capable of playing back the synthesis of the pitch values from the Objects list and the PitchEditor window.) Unreliable measures can be corrected within Praat's constraints on pitch editing, or removed.

One convention has been to then take the lowest pitch value in a corrected pitch trace as the speaker's baseline, and the highest pitch value as the speaker's topline. The median pitch value (the pitch value at the 50th percentile of the distribution of all values) can be taken as the middle of the speaker's range. The median generally provides a more appropriate estimate of the middle of the speaker's range then the mean (the sum of all pitch values divided by the number of values): there are usually more pitch values towards the bottom of the speaker's range than the top (Lennes et al., 2015; this is also shown in results in Gorisch et al., 2012, Horii, 1975).

Another method involves generating pitch values for all available speech from that speaker (Gorisch, Wells, & Brown, 2012; Lennes, Stevanovic, Alto, & Palo, 2015). However, even with adjustment to pitch detection parameters it is almost inevitable that there will be erroneous pitch values due to a number of factors including background noise, overlapping talk and changes in phonation.[4] These erroneous values tend to occur at very high and very low frequencies, so while their impact on measures of the middle of the range may be negligible, they may have a significant impact on estimates of the top and bottom of the range.

When presenting and interpreting pitch traces researchers and readers should also be aware of the interaction between frequency and intensity (Fletcher, 1934, Wier, Jesteadt, and Green, 1977). The interaction of frequency and intensity in auditory perception is not represented visually in conventional pitch traces where all pitch

[4] Non-modal phonation can produce very low frequencies (e.g. creak) and very high ones (e.g. falsetto). A decision therefore needs to be made as to whether these measures are included in estimates of range. Their inclusion could radically alter the visual representation of the speaker's range. Since speakers usually only produce talk in those registers infrequently, a practical solution is to exclude such stretches from the estimation of the speakers' pitch range (hence locutions such as "normal speaking range") and then adjust individual pitch traces when necessary to show non-modal portions outside the normal pitch range (see for example G. Walker, 2016, figure 5a, figure 6).

values are represented by an identical dot, or a line through the values. G. Walker (2012), published in *Language and Speech*, attempts to take the interaction of frequency and intensity into account by plotting pitch values in different shades of gray depending on the intensity of the signal. Another way to convey information about intensity is to present an intensity trace. Kohler (2013), published in *Phonetica*, presents composite figures which include a waveform, spectrogram, pitch trace and an intensity trace. Figure 20 is a visual representation of the same utterance as Figure 1 of Kohler (2013), drawn to take into account some of the suggestions made in this article.

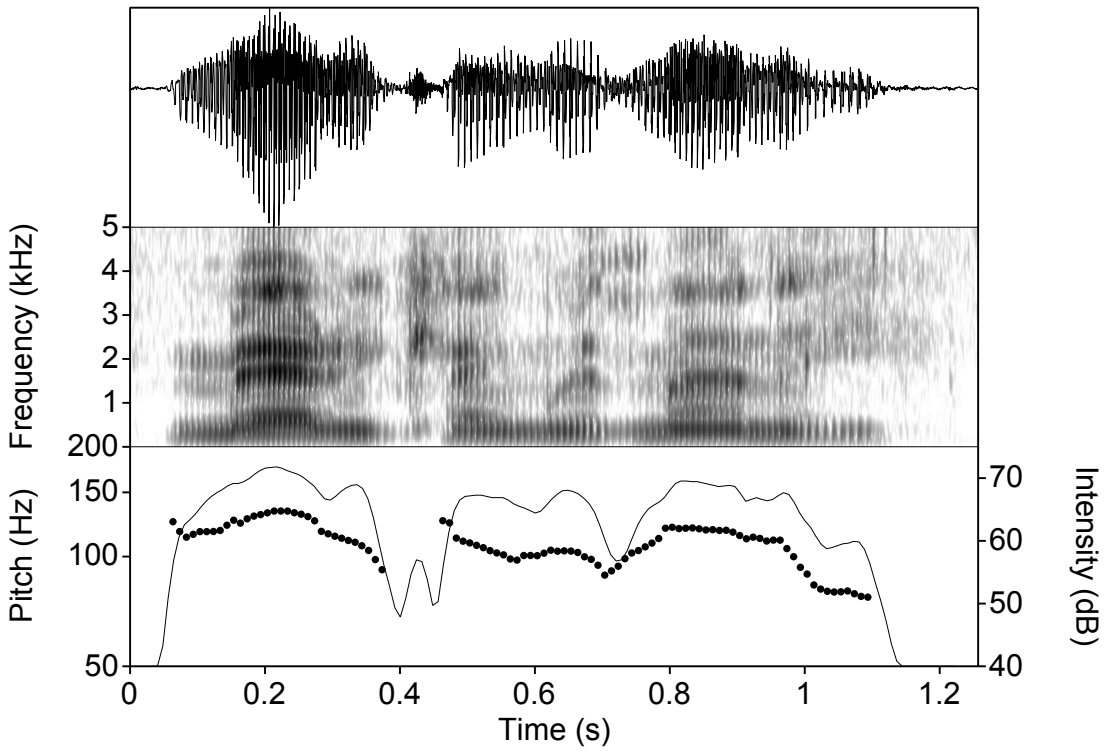


Figure 20 Waveform, spectrogram, pitch trace (dotted) and intensity trace (dotted) of an adult male producing "Mary came with Manny"

4.3 Adjusting and presenting spectrograms

Source K and source U are the only sources to include spectrograms as part of their visual representations. They are not discussed at all in source K and receive scant attention in source U with only one (on p. 399) being discussed in any detail. However, that spectrogram shows rather too much material to identify the relevant features. It is straightforward to address this by allowing material to take up more horizontal space. Figure 21 shows a spectrogram of “key yih” from the same utterance (“A LO:TTA TURKEY YIHKNOW I DON’T HAVE ONE:”). The spectrogram was drawn with a dynamic range of 35 dB. (Praat’s default display in the SoundEditor window is 70 dB; the default when drawing from the Objects list is 50 dB.) Decreasing the dynamic range of a spectrogram decreases the range of values shown as shades of grey which may visually enhance relevant features; conversely, increasing the dynamic range increases the range of values in the spectrogram shown as shades of grey which will make the spectrogram look darker and may

obscure features. The window length of the spectrogram has been reduced from Praat's default of 0.005 s to 0.003 s to take into account the high pitch of the speaker's voice at this point. As in source U, the frequency range is from 0 Hz to 5000 Hz which in the absence of a specific need to give greater clarity to lower frequencies (see Figure 4) or to show higher frequencies, is generally appropriate when presenting spectrograms.

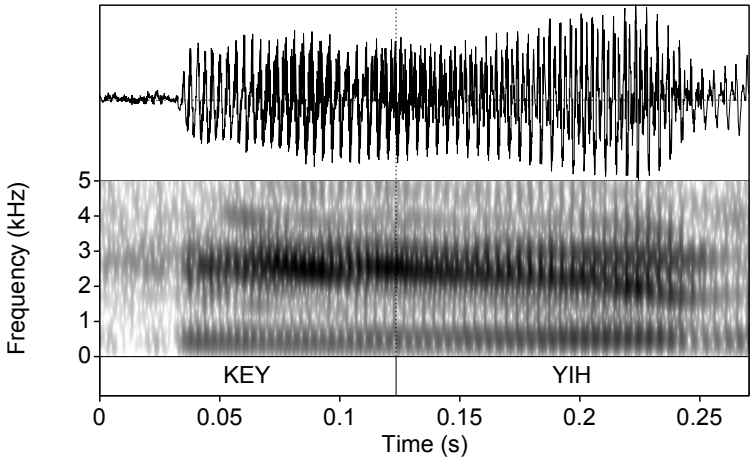


Figure 21 Labelled spectrogram of an adult female's speech

Drawn in this way it is straightforward for the reader to see the striations in the spectrogram corresponding to voicing. The inclusion of the waveform supports the claim in source U of "no break in voicing" (source U, p. 400): the waveform is periodic across the join. It is also straightforward for the reader to see from the spectrogram that the vocal tract resonances stay relatively steady across the join between the two words: note the dark bands centered around 0.5 kHz and 2.5 kHz. This reflects what source U refer to as the "single palatal place of articulation" (source U, p. 400). The mode of presentation in Figure 21 is similar to that adopted by Local and Walker (2012) in their study of turn- and talk-projecting phonetic features published in the *Journal of the International Phonetic Association.* Figure 22 is a visual representation of the same utterance as Figure 3a of Local and Walker (2013), drawn to take into account some of the suggestions made in this article.

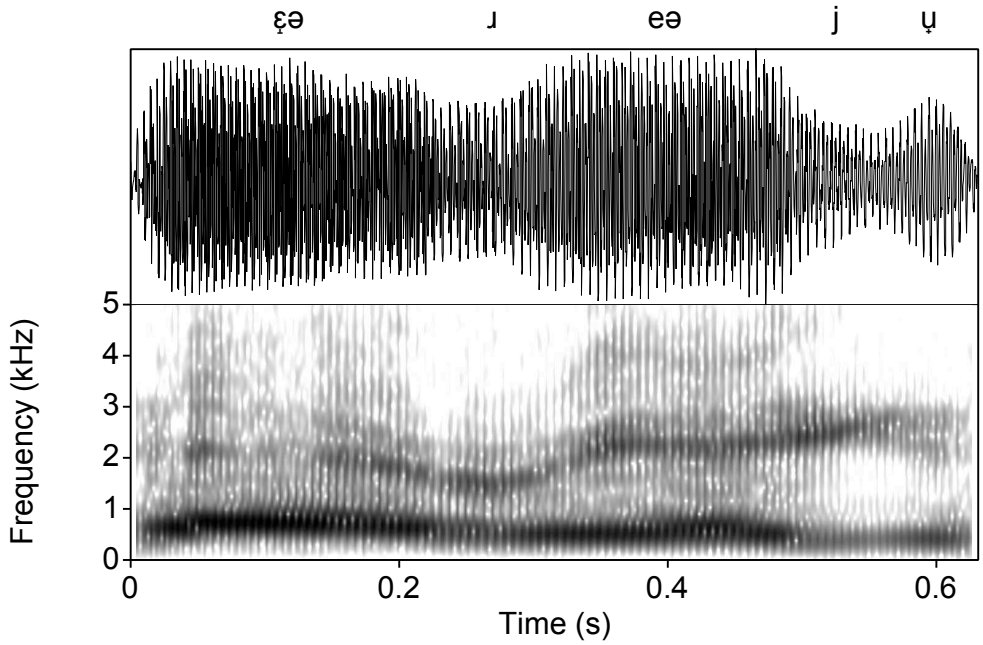


Figure 22 Spectrogram and waveform of an adult female producing “area you”; labels at the top of the figure are aligned with the relevant portion of the spectrogram and waveform and are rendered in the notation of the International Phonetic Association

5 Summary and conclusions

The primary means for conveying information about what a researcher can hear in recordings of talk-in-interaction is what the reader can see: descriptions, transcriptions, numerical measures and visual representations of acoustic data. This article has surveyed the use of visual representations of acoustic data in ROLSI and made some suggestions as to how they might be prepared and used more effectively. Comparisons have been made with relevant visual representations of acoustic data prepared by expert phoneticians and published in other journals. This has shown some ways visual representations could have been more effectively prepared and used. Visual representations of acoustic data are an important resource a researcher can use to allow the reader to ‘get at’ the data. They become all the more important when the reader does not have access to the original recordings. The main message is not that visual representations should be avoided, but that more care needs to be taken over their preparation and use: as much care as is taken over the description and analysis of turn and sequence. The focus here has been on the visual representations themselves rather than the analytic procedures researchers employ when dealing with data. However, it is possible that shortcomings in the preparation and use of visual representations of acoustic data reflect an incomplete understanding of relevant technical issues.

The main findings of the survey can be summarised as follows:

- visual representations of acoustic data are becoming more common
- visual representations are not always prepared or used in such a way to allow them to reach their full potential, in terms of providing corroborative evidence for researchers’ claims and/or allowing readers to independently verify the claims being made

- phonetic features of relevance to the researcher's argument may not receive adequate visual representation

Some of the preceding discussion has been concerned with providing advice on the effective preparation and use of visual representations. Those suggestions can be summarised as follows:

The researcher(s) should ensure that

- the selected visual representations are those best suited to the analytic task
- visual representations of acoustic data can be fully and confidently interpreted by the reader
- as far as possible, relevant perceptual factors are taken into account when preparing visual representations
- visual representations are integrated into the discussion, and discussed accurately
- space is used efficiently in visual representations
- irrelevant information in visual representations is minimised, and relevant information maximised

Shortcomings in the preparation and use of visual representation of acoustic data in CA/IL research are not a rare event: in all sources the visual representations could have been prepared and used with more care. This would have helped ensure corroborative evidence for the researchers' claims was provided and that those claims could be independently verified by the reader. Shortcomings in the preparation and use of visual representation of acoustic data in CA/IL research are not unique to articles published in ROLSI.

Scripts

Praat scripts for creating visual representations including the original figures in this article are available via the author's homepage: http://gareth-walker.staff.shef.ac.uk/praat/visreps/.

References

Baken, R. & Orlikoff, R. F. (2000). *Clinical measurement of speech and voice* (2nd ed.). San Diego: Singular.

Barthel, H. & Quené, H. (2015). Acoustic-phonetic properties of smiling revised -- measurements on a natural video corpus. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK. Retrieved from http://icphs2015.info/Proceedings.aspx

Barth-Weingarten, D. (2011). Double sayings of German *JA*--more observations on their phonetic form and alignment function. *Research on Language and Social Interaction, 44*(2), 157–185. doi:10.1080/08351813.2011.567099

Betz, E. & Golato, A. (2008). Remembering relevant information and withholding relevant next actions: the German token *achja*. *Research on Language and Social Interaction, 41*(1), 58–98. doi:10.1080/08351810701691164

Boersma, P. & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Retrieved from http://www.fon.hum.uva.nl/praat/

Clayman, S. E. & Raymond, C. W. (2015). Modular pivots: a resource for extending turns at talk. *Research on Language and Social Interaction, 48*(4), 388–405. doi:10.1080/08351813.2015.1090112

Couper-Kuhlen, E. (1996). The prosody of repetition: on quoting and mimicry. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation: interactional studies* (pp. 366–405). Cambridge: Cambridge University Press.

Fletcher, H. (1934). Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure. *Journal of the Acoustical Society of America, 6*(2), 59–69. doi:10.1121/1.1915704

Golato, A. (2012). German *oh*: marking an emotional change of state. *Research on Language and Social Interaction, 45*(3), 245–268. doi:10.1080/08351813.2012.699253

Golato, A. & Fagyal, Z. (2008). Comparing single and double sayings of the German response token ja and the role of prosody: a conversation analytic perspective. *Research on Language and Social Interaction, 41*(3), 241–270. doi:10.1080/08351810802237834

Gorisch, J., Wells, B., & Brown, G. J. (2012). Pitch contour matching and interactional alignment across turns: an acoustic investigation. *Language and Speech, 55*(1), 57–76. doi:10.1177/0023830911428874

Helasvuo, M. L., Laakso, M., & Sorjonen, M. L. (2004). Searching for words: syntactic and sequential construction of word search in conversations of Finnish speakers with aphasia. *Research on Language and Social Interaction, 37*(1), 1–37. doi:10.1207/s15327973rlsi3701_1

Hellermann, J. (2005). Syntactic and prosodic practices for cohesion in series of three-part sequences in classroom talk. *Research on Language and Social Interaction, 38*(1), 105–130. doi:10.1207/s15327973rlsi3801_4

Hepburn, A. (2004). Crying: notes on description, transcription, and interaction. *Research on Language and Social Interaction, 37*(3), 251–290. doi:10.1207/s15327973rlsi3703_1

Hepburn, A. & Potter, J. (2007). Crying receipts: time, empathy, and institutional practice. *Research on Language and Social Interaction, 40*(1), 89–116. doi:10.1080/08351810701331299

Heritage, J. (2012). Epistemics in action: action formation and territories of knowledge. *Research on Language and Social Interaction, 45*(1), 1–29. doi:10.1080/08351813.2012.646684

Hollien, H., Hollien, P. A., & de Jong, G. (1997). Effects of three parameters on speaking fundamental frequency. *Journal of the Acoustical Society of America, 102*(5), 2984–2992. doi:10.1121/1.420353

Horii, Y. (1975). Some statistical characteristics of voice fundamental frequency. *Journal of Speech, Language, and Hearing Research, 18*(1), 192–201.

doi:10.1121/1.1981923

Innes, B. (2007). “Everything happened so quickly?” - HRT intonation in New Zealand courtrooms. *Research on Language and Social Interaction, 40*(2-3), 227–254. doi:10.1080/08351810701354672

Johnson, K. (2003). *Acoustic & auditory phonetics* (2nd ed.). Oxford: Blackwell.

Kaimaki, M. (2011). Transition relevance and the phonetic design of English call openings. *Journal of Pragmatics, 43*(8), 2130–2147. doi:10.1016/j.pragma.2011.01.008

Kohler, K. J. (2008). ‘Speech-smile’, ‘speech-laugh’, ‘laughter’ and their sequencing in dialogic interaction. *Phonetica, 65*(1-2), 1–18. doi:0.1159/000130013

Kohler, K. J. (2013). From communicative functions to prosodic forms. *Phonetica, 70*(1-2), 24–65. doi:10.1159/000351415

Kreiman, J. & Sidtis, D. (2011). *Foundations of voice studies: an interdisciplinary approach to voice production and perception*. Malden, MA: Blackwell Publishing Ltd.

Lennes, M., Stevanovic, M., Alto, D., & Palo, P. (2015). Comparing pitch distributions using Praat and R. *The Phonetician, 111-112*, 35–53.

Local, J. (2005). On the interactional and phonetic design of collaborative completions. In W. Hardcastle & J. M. Beck (Eds.), *A figure of speech: a festschrift for John Laver* (pp. 263–282). Mahwah, New Jersey: Lawrence Erlbaum.

Local, J. & Walker, G. (2012). How phonetic features project more talk. *Journal of the International Phonetic Association, 42*(3), 255–280. doi:10.1017/s0025100312000187

MacMartin, C., Coe, J. B., & Adams, C. L. (2014). Treating distressed animals as participants: I know responses in veterinarians’ pet-directed talk. *Research on Language and Social Interaction, 47*(2), 151–174. doi:10.1080/08351813.2014.900219

Mitchell, R. W. (2001). Americans’ talk to dogs: similarities and differences with talk to infants. *Research on Language and Social Interaction, 34*(2), 183–210. doi:10.1207/S15327973RLSI34-2_2

Nakamura, K. (2001). Gender and language in Japanese preschool children. *Research on Language and Social Interaction, 34*(1), 15–43. doi:10.1207/S15327973RLSI3401_2

Nishio, M. & Niimi, S. (2008). Changes in speaking fundamental frequency characteristics with aging. *Folia Phoniatrica et Logopaedica, 60*(3), 120–127. doi:10.1159/000118510

Ogden, R. (2001). Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association, 31*(1), 139–152. doi:10.1017/s0025100301001116

Ogden, R. (2009). *An introduction to English phonetics.* Edinburgh: Edinburgh University Press.

Ogden, R. (2013). Clicks and percussives in English conversation. *Journal of the International Phonetic Association, 43*(3), 299–320. doi:10.1017/s0025100313000224

Peppé, S. (2009). Why is prosody in speech-language pathology so difficult? *International Journal of Speech-Language Pathology*, 11(4), 258–271.

Persson, R. (2013). Intonation and sequential organization: formulations in French talk-in-interaction. *Journal of Pragmatics, 57*, 19–38. doi:10.1016/j.pragma.2013.07.004

Pierce, J. R. (1992). *The science of musical sounds*. Revised edition. New York: W.H. Freeman and Company.

Pillet-Shore, D. (2012). Greeting: displaying stance through prosodic recipient design. *Research on Language and Social Interaction, 45*(4), 375–398. doi:10.1080/08351813.2012.724994

Podesva, R. J., Callier, P., Voigt, R., & Jurafsky, D. (2015). The connection between smiling and GOAT fronting: embodied affect in sociophonetic variation. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK. Retrieved from http://icphs2015.info/Proceedings.aspx

Reetz, H. & Jongman, A. (2009). *Phonetics: transcription, production, acoustics, and perception*. Oxford: Wiley-Blackwell.

Rendle-Short, J. (2005). Managing the transitions between talk and silence in the academic monologue. *Research on Language and Social Interaction, 38*(2), 179–218. doi:10.1207/s15327973rlsi3802_3

Robinson, J. D. & Kevoe-Feldman, H. (2010). Using full repeats to initiate repair on others' questions. *Research on Language and Social Interaction, 43*(3), 232–259. doi:10.1080/08351813.2010.497990

Stivers, T. & Sidnell, J. (2016). Proposals for activity collaboration. *Research on Language and Social Interaction, 49*(2), 148–166. doi:10.1080/08351813.2016.1164409

Szczepek Reed, B. (2015). Managing the boundary between "yes" and "but": two ways of disaffiliating with German *ja aber* and *jaber*. *Research on Language and Social Interaction, 48*(1), 32–57. doi:10.1080/08351813.2015.993843

Szczepek Reed, B. & Persson, R. (2016). How speakers of different languages extend their turns: word linking and glottalization in French and German. *Research on Language and Social Interaction, 49*(2), 128–147. doi:10.1080/08351813.2016.1164405

Tartter, V. C. (1980). Happy talk: perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics, 27*(1), 24–27. doi:10.3758/bf03199901

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

Walker, G. (2012). Coordination and interpretation of vocal and visible resources: 'Trail-off' conjunctions. *Language and Speech, 55*(1), 141–163. doi:10.1177/0023830911428858

Walker, G. (2013). Phonetics and prosody in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 455–474). Oxford: Wiley-Blackwell.

Walker, G. (2016). Phonetic variation and interactional contingencies in simultaneous responses. *Discourse Processes, 53*(4), 298–324. doi:10.1080/0163853x.2015.1056073

Walker, T. (2014a). Form ≠ function: the independence of prosody and action. *Research on Language and Social Interaction, 47*(1), 1–16. doi:10.1080/08351813.2014.871792

Walker, T. (2014b). The independence of phonetic form and interactional accomplishments. *Research on Language and Social Interaction, 47*(1), 23–27. doi:10.1080/08351813.2014.871796

Wier, C. C., Jesteadt, W., & Green, D. M. (1977). Frequency discrimination as a function of frequency and sensation level. *Journal of the Acoustical Society of America, 61*(1), 178–184. doi:10.1121/1.381251

Wiggins, S. (2002). Talking with your mouth full: gustatory *mmms* and the embodiment of pleasure. *Research on Language and Social Interaction, 35*(3), 311–336. doi:10.1207/S15327973RLSI3503_3

Wilkinson, R., Beeke, S., & Maxim, J. (2010). Formulating actions and events with limited linguistic resources: enactment and iconicity in agrammatic aphasic talk. *Research on Language and Social Interaction, 43*(1), 57–84. doi:10.1080/08351810903471506